УДК 004.912

О.В. Бармак[1], Е.А. Манзюк[1], Ю.В. Крак[2,3] , А.І.Куляс[3]
[1]Хмельницький національний університет, Україна
вул. Інститутська, 11, м. Хмельницький, 29000
[2] Київський національний університет імені Тараса Шевченка, Україна
пр-т Глушкова, 4д, м. Київ, 03680
[3] Інститут кібернетики імені В.М. Глушкова НАН України, Україна
пр-т Глушкова, 40, м. Київ, 03187

ПРИНЦИПИ ТА ПІДХОДИ ДО ФОРМУВАННЯ АНСАМБЛІВ КЛАСИФІКАТОРІВ НА ПІДСТАВІ АГРЕГАТУВАННЯ ЇХ РЕЗУЛЬТАТІВ

O.V.Barmak[1] , E.A.Manziuk[1] , Iu.V.Krak[2,3], A.I.Kulias[3]
[1] Khmelnytsky National University, Ukraine
11, Instytutska St., Khmelnytsky, 29000
[2]Taras Shevchenko National University of Kyiv, Ukraine
4d, Hlushkov ave., Kyiv, 03680
[3] V.M.Hlushkov Institute of Cybernetics, Ukraine
40, Hlushkov ave., Kyiv, 03187

PRINCIPLES AND APPROACHES TO THE FORMATION OF ANSAMBLES BASED ON ANALYSIS RESULTS

У статті розглянуто та проаналізовано принципи формування та застосування ансамблів класифікаторів. Розглянуто переваги використання підходів агрегатування рішень класифікаторів. Аналіз принципів перерозподілу даних при розширенні ансамблю дав змогу сформувати вимоги до класифікаторів ансамблю, на підставі чого було виявлено умови вибору класифікаторів та критерії їх достатності. На базі критеріїв достатності встановлено умови застосовності принципу агрегації для граничних множин. Проведено аналіз неоднозначності прийняття рішень для симетричних множин класифікаторів. Визначено умови формування множини класифікаторів та встановлені критерії її достатності для вирішення завдання класифікації. Встановлено, що подальше розширення множини класифікаторів понад критерії достатності привносить помилки класифікації до множини правильних рішень усіх класифікаторів. Розширення множини класифікаторів дозволяє сформувати набір сполучень, який є трикутником Паскаля та проаналізувати граничний перерозподіл даних у процесі збільшення ансамблю.

Ключові слова: ансамблі, трикутник Паскаля, теорема Кондорсе, класифікація, and-gate, голосування простою більшістю

The article considers principles of the formation and application of classifiers' ensembles. The advantages of using the approaches of aggregation of classifier solutions are considered. An analysis of the principles of data redistribution with the extended ensemble made it possible to formulate requirements for the classifiers of the ensemble. On the basis of what was found the conditions for the selection of classifiers and the criteria for their adequacy. On the basis of the sufficiency criteria, the applicability conditions for the principle of aggregation for boundary sets were established. An analysis of the ambiguity of decision making for symmetric sets of classifiers is carried out. The conditions for forming a set of classifiers and the criteria for its adequacy for solving the recognition problem are determined. It is established that further expansion of the set of classifiers over the sufficiency criterion brings classification errors to the set of correct solutions of all classifiers. The extension of the classifier set allows us to form a set of connections that is a Pascal triangle and to analyze the marginal redistribution of data in the process of increasing the ensemble.

Keywords: ensembles, Pascal triangle, Condorcet's theorem, classification, and-gate, majority voting

Вступ

На сьогоднішній день існує досить широкий набір засобів інтелектуального аналізу даних: методи статистичного аналізу, нейронні мережі, SVM тощо. Таке різноманіття засобів свідчить про відсутність універсальних підходів для вирішення прикладних задач.

Застосування різноманітних методів та засобів моделювання до одного і того ж набору даних може мати різну мету та різні площини дослідження. Це може бути спрощена модель, результати якої

досить легко інтерпретувати, однак точність її буде досить невисокою. Іншим випадком може бути побудова складної моделі, однак результати її роботи досить важно інтерпретувати. Загалом складні моделі, що дають найбільш точні результати, у переважній більшості складно інтерпретуються.

Як результат, для практичного застосування методів інтелектуального аналізу даних будують моделі, виходячи з обраних цілей та знаходження компромісу між такими показниками як точність, складність, інтерпретованість.

Досить значна частина досліджень направлена на розробку моделей, які дають більш точні результати, зважаючи на те, що для кінцевих користувачів питання прозорості результатів є неважливим та досить суб'єктивним. Проте інтерпретованість в певних областях застосування машинного навчання є особливо важливим показником.

Точність роботи моделей залежить від якості даних у проекції предметної області та методів аналізу. Так як різноманіття моделей у площині машинного навчання досить широке, природним є поєднання декількох підходів та створення ансамблів моделей. Це дозволяє використати їх переваги та підвищити якість рішення.

Постановка проблеми

Набір моделей агрегатованого класифікатора навчають та отримують результати прогнозування, які надалі поєднують і на базі поєднання прогнозувань отримують певне рішення ансамблю.

Кінцева кількість моделей навчається на загальній множині прецедентів або її підмножині та одержують результати прогнозу. Однак виникає питання підходів до агрегації результатів прогнозування окремих моделей та формування гіпотези про стан системи. Оцінка гіпотез та їх агрегатований вибір залежить від багатьох факторів. Ці фактори пов'язані з характером розподілу даних та їх параметрів, зв'язків між параметрами, цільовими функціями на площині параметрів тощо. Окрема модель, яку навчили на множині прецедентів, може не містити властивості узагальнення та не показати хорошу точність на тестовій вибірці. Однак із застосуванням ансамблю, та використавши достатню кількість моделей, які навчаються на одній і тій же множині, можемо на підставі багатьох результатів зменшити випадковість результуючого рішення шляхом комбінування множини результатів. Це призведе до компенсації випадкових факторів недосконалості окремих моделей та підвищить кінцеву точність агрегації. Такий результат можливо отримати у випадку взаємної компенсації недоліків моделей. У цьому випадку випадковість факторів дозволяє підвищити точність ансамблю. Моделі врівноважують випадковість результатів один одного та збалансовують спільне рішення, яке найбільш правдоподібне до цільової функції. Це дозволяє підвищити результат та взаємно компенсувати помилки кожної окремої моделі. Вважається, що середня оцінка рішень множини моделей на базі незалежних навчальних множин завжди зменшує значення середньоквадратичної помилки.

Зменшення помилки при використанні ансамблю базується на такому явищі, що має назву ефект геніальності натовпу. Однак тут необхідно також враховувати як розподіл результатів моделей, так і факторів параметрів даних.

Агрегатування результатів рішень класифікаторів

Для отримання результуючого рішення із рішень, що були отримані кожною моделлю, як правило, використовують такі підходи:

- голосування: результатом спільного рішення є проста більшість із загальної множини моделей;
- зважене голосування: кожному голосу окремої моделі надається певна вага за обраними критеріями ансамблю;
- середнє значення: результати всього ансамблю визначаються як просте

середнє значення результатів кожної моделі, за наявності ваг результати збільшуються на величину ваги.

Побудова рішень за голосуваннями в окремих випадках може змінювати та враховувати специфіку мети голосування. Агрегатування результатів може бути не тільки за більшістю, але і набувати інших видів. Як приклад, можна навести такі підходи до поєднання результатів за правилами:

- and-gate: всі моделі з множини прийняли однакове рішення;
- or-gate: хоча би одна модель прийняла рішення; як приклад, можна навести приналежність інформації до певного класу у бінарній класифікації, тобто усі моделі прийняли рішення про неналежність до класу, а одна модель прийняла рішення що все ж таки належить до нього, то у цьому випадку приймається спільне рішення ансамблю, що інформація належить до цього класу;
- k-out-of-N: у випадку, коли k моделей з N прийняли рішення. У цьому випадку число N не завжди відповідає загальній кількості моделей в ансамблі;
- majority vote: якщо більшість моделей прийняла рішення, тоді спільне рішення, тобто рішення ансамблю, відповідає цьому рішенню.

Метод ансамблів загалом дозволяє отримати наступні переваги:

1. Мінімізація впливу випадковостей. Оскільки об'єднуючий класифікатор компенсує помилки кожного з базових класифікаторів, тобто середня помилка спільного рішення зменшується, то це призводить до зменшення випадковостей на результуюче рішення.
2. Зменшення дисперсії. Спільне рішення множини моделей загалом дозволяє отримати кращу оцінку, ніж рішення окремої моделі. Ймовірність того, що загальна сукупність моделей знайде правильне рішення значно вища. Загалом, оптимальне рішення площини рішень в умовах ансамблю може мати глобальний оптимум. Це випливає з того, що пошук рішення здійснюється множиною моделей з різних точок цього простору. Таким чином, якоюсь моделлю це рішення буде знайдене.
3. Репрезентативність. Існує імовірність того, що множина гіпотез моделей не містить найкращої гіпотези. У такому випадку є можливість доповнення множини гіпотез комбінованими гіпотезами. Таким чином, використання ансамблів дозволяє розширити множину можливих гіпотез. Це дозволяє підвищити імовірність знаходження оптимальної гіпотези та отримати результат за межами множини базових гіпотез моделей, тобто отримати нові можливості на базі існуючих.

Таким чином, обираємо сукупність моделей, поєднуємо їх за обраним правилом та отримуємо результат агрегатування. Це дає нам такі переваги:

- отримання більш складної моделі, ніж будь-яка, що є частиною ансамблю;
- підвищення точності;
- уникнення перенавчання та недонавчання;
- можливість поєднання ознак різної природи.

Граничні випадки агрегатування рішень класифікаторів

Вважається, що використання агрегацій класифікаторів виходить з теореми Кондорсе про журі присяжних [1], згідно з якою кожний член присяжних має незалежну думку і рішення не залежить від інших та має однакову ймовірність виявитись правильним p. Якщо ймовірність винесення членом журі правильного рішення знаходиться в межах $0.5 < p < 1$, тоді ймовірність винесення журі правильного рішення збільшується зі збільшенням кількості присяжних та наближається до одиниці. У випадку $0 < p < 0.5$ імовірність правильного рішення журі зменшується та набли-

жається до нуля зі збільшенням кількості присяжних.

У такому випадку ймовірність того, що журі винесе правильне рішення згідно з правилом простої більшості буде мати вигляд:

$$P_n = \sum_{i=\frac{n}{2}+1}^{n} \frac{n!}{(n-i)!i!} p^i (1-p)^{(n-i)}, \quad (1)$$

де n – кількість присяжних (приймається парна кількість);

p – імовірність правильного рішення присяжного;

$\frac{n}{2}+1$ – мінімальна більшість присяжних журі.

Теорема базується на припущеннях [1]:

1. Необхідність загальної ймовірності правильних рішень у всіх присяжних;
2. Кожний член журі повинен робити вибір, що не залежить від вибору інших присяжних;
3. Кожен присяжний голосує справедливо, враховуючи лише своє власне судження щодо правильного рішення.

Якщо кожен присяжний має свою власну ймовірність рішення, то це не призводить до зміни теореми.

При повторі рішення одного із членів, імовірність не вища ймовірності прийняття рішення будь-кого з присяжних. Невелика кореляція не призводить до неможливості прийняття загального правильного рішення.

Третя вимога говорить про джерело інформації, на базі якого присяжні приймають рішення.

Граничний випадок використання ансамблів з простим голосуванням. При задачі бінарної класифікації, де $\{f_i(x)\}_{i=1}^{M}$ – бінарні класифікатори. Для випадкових величин, які задаються відповідним ймовірнісним простором $\xi_i = I[f_i(x)] \sim Ber(p)$, $p(0,1)$. При цьому $E[\xi_i] = p$, $D[\xi_i] = p(1-p)$. Необхідною умовою є незалежність класифікаторів, тоді $\{\xi_i\}_{i=1}^{M}$ незалежні. Використаємо $\bar{\xi} = \frac{1}{M}\sum_{i=1}^{M}\xi_i$ та $E[\bar{\xi}] = p$, $D[\bar{\xi}] = \frac{p(1-p)}{M}$. Так як ансамбль $F(x)$, який складається з $\{f(x)_1,...,f(x)_M\}$, голосуванням більшістю голосів неправильно класифікує за умови $\bar{\xi} > 0.5$.

Імовірність помилки

$$P(F(x)) = \sum_{i=\frac{M}{2}+1}^{M} \frac{M!}{(M-i)!i!} p^i (1-p)^{(M-i)} \quad (2)$$

Для граничного випадку, за умови $M \to \infty$ значення $D[\bar{\xi}] \to 0$, розподіл величини $\bar{\xi}$ вироджується. При збільшені величини M імовірність помилки ансамблю $P(F(x))$ наближається до нуля, якщо $p < 0.5$. На підставі цього, при побудові ансамблю простим голосуванням, говорять про можливість побудови ідеального класифікатора.

Рішення ансамблю буде мати підвищену точність, оскільки неправильні рішення класифікатора компенсуються відповідними правильними рішеннями інших класифікаторів. Щоб отримати очікуваний результат, класифікатори повинні бути різноманітними. Різноманітність полягає в тому, що вони повинні компенсувати неправильні рішення, а не поділяти помилки [2, 3]. Окрім різноманітності, для того щоб випадково обрані класифікатори в ансамблі досягли високої точності, вони повинні бути самостійно слабкими, проте бути достатньо точними та вказувати точність, яка щонайменше перевищує 50% [4].

Однак відомі дослідження [5], у яких автори наводять приклад про те, що теорема Кондорсе про журі присяжних часто цитується некоректно, як основа твердження про те, що практичне застосування призводить до покращення класифікації, тобто зменшення ймовірності помилки за умови, коли класифікатори є достатньо ефективними та некоре-

льованими. Автори зосереджують увагу на тому, що вимоги застосування повинні бути поставлені для незалежних (а не просто некорельованих) помилок класифікації (а не самих класифікаторів). Наведений приклад демонструє, що найбільш часті приклади інтерпретації вимог застосування є помилковими, зокрема для голосування простою більшістю.

Існують випадки, коли використання голосування простою більшістю необхідно застосувати в умовах умовно незалежних класифікаторів [6]. Проведені дослідження на великій кількості прикладів виявили той факт, що, як правило, такий вид агрегації працює досить добре. Проте також було виявлено, що умовна незалежність не є достатньою умовою для забезпечення того, щоб голосування більшістю голосів завжди забезпечувало кращу класифікаційну ефективність, ніж окремі класифікатори.

Таким чином, хоча в значній кількості випадків агрегатування за голосуванням переважною більшістю голосів дає хороші результати, існують випадки, коли застосування загальноприйнятих вимог до використання таких видів агрегатування не є ефективним. Крім того, трактування причин наявності таких явищ викликає дискусію у їх формуванні.

Міграція даних при формуванні набору сполучень рішень класифікаторів

У зв'язку з цим розглянемо поводження та взаємодію моделей у їхній сукупності до формування принців та правил агрегатування результатів в ансамблі. Крім того, визначимо поведінку та принципи формування сукупності класифікаторів на підставі їх поводження та результатів класифікації кожного окремо взятого класифікатора.

Базуючись на дослідженнях [7], застосуємо універсальну навчальну множину та проведемо класифікацію на тестовій вибірці класифікаторами із формуванням узагальнюючих результатів у системі їх поведінки.

За результатами класифікації моделі формують множини правильно класифікованих даних. Причому існує певна сукупність даних, які правильно класифікували декілька моделей. Ці дані певним чином перерозподіляються із додаванням нових класифікаторів до множини застосовних.

Сформуємо підмножини з універсальної множини даних за усіма можливими комбінаціями спільних правильних рішень моделей на площині тестових даних. Розглянемо бінарну класифікацію правильних рішень $\{0,1\}$. Тобто, якщо рішення моделі правильне, то позначимо 1, якщо ні – 0.

Для прикладу, з трьох класифікаторів одержимо такі сполучення $\{1,1,1\}$, $\{1,1,0\}$, $\{1,0,0\}$, $\{0,0,0\}$ з числами сполучень $\binom{3}{3}$, $\binom{3}{2}$, $\binom{3}{1}$, $\binom{3}{0}$. Як відомо, кількість сполучень n по k рівна біноміальному коефіцієнту. Якщо будемо додавати нову модель, отримаємо набори сполучень, які формують трикутник Паскаля.

Зобразимо відповідність між трикутником Паскаля та діаграмою Венна (рис.1).

Граничні набори, таким чином, формують множину даних, які були правильно класифіковані усіма моделями з множини моделей ансамблю, та множину даних, які усі моделі класифікували неправильно. У процесів розширення множини ансамблю відбувається процес перерозподілу множин даних та міграція даних у сусідні множини. Таким чином, внесення нової моделі до ансамблю вносить зміни у перерозподіл даних у множинах.

Як відомо, множина ансамблю є скінченною та повинна формуватися з множини можливих моделей, вносячи певну інформаційну цінність в ансамбль. Модель повинна доповнювати ансамбль та надавати йому додаткові властивості із класифікації. Тобто, внесення моделі в ансамбль має сенс тільки у тому випадку коли вона правильно класифікувала дані, які не були правильно класифіковані жодною з моделей, що входять до ансам-

блю. Практично це визначається переходом даних зі сполучення $\{0{,}0{,}0\}$ в $\{1{,}0{,}0\}$. Крім того, вибір моделей з усіх можливих повинен здійснюватися за ознакою максимізації кількості даних згідно з зазначеним переходом.

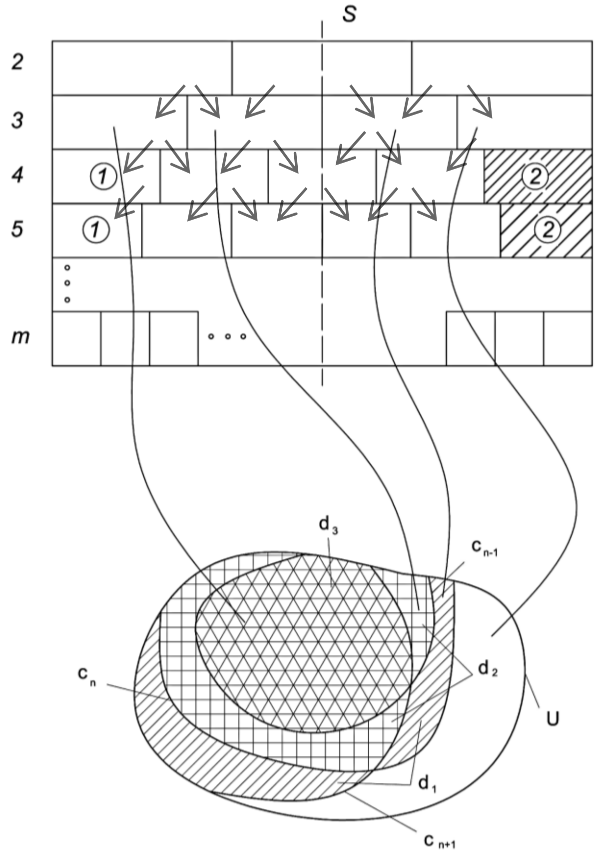


Рис. 1. Розподіл даних із внесенням додаткових моделей в множину застосовних

Така форма додавання моделей до ансамблю призводить до поступового зменшення множини сполучення $\{0{,}0{,}0\}$.

Тобто, множина даних, які не були правильно розпізнані жодним класифікатором з ансамблю вироджується на підставі переходу

$$\bigcup_{C} C \setminus U = \{\ \}. \tag{3}$$

На рис. 1 це є заштрихована зона 2, в якій відсутні дані.

Так як усі дані були розпізнані хоча би одним класифікатором з ансамблю, подальше розширення ансамблю не має значення. Це говорить про те, що дані можуть бути розпізнані в межах поставленої задачі.

Таким чином, можливо записати умови формування ансамблю класифікаторів з множини можливих W.

$$C = \{c \mid A(c)\}. \tag{4}$$

$$\underset{i\in[1,\ldots,|C|]}{A(c_i)} = \begin{cases} c_i = \underset{j\in[i,\ldots,|W|]}{\arg\max}\, b_j,\ b_j = \begin{cases} \bigcap_{k=1}^{i} C_k; \\ \bigcup_{k=1}^{i} C_k \setminus U. \end{cases} \\ \bigcup_{C} C \setminus U \neq \{\ \}. \end{cases} \tag{5}$$

Подальше додавання моделей до ансамблю призводить тільки до перерозподілу даних та зменшення множини сполучення $\{1{,}1{,}1\}$. Множина, яка формується з правильних рішень усіх моделей зменшується, таким чином, подальше додавання моделей не привносить до ансамблю необхідної інформаційної цінності та вносить помилки до множини сполучення $\{1{,}1{,}1\}$.

Агрегатування моделей ансамблю

Сформований набір сполучень має лінію симетрії S (рис. 1), яка проходить посередині розподілу сполучень та формує вісь невизначеності при формуванні принципів агрегатування рішень моделей.

Загалом немає можливості визначити принципи та підходи, які дозволили б встановити сукупність даних, для яких є однозначною характеристика по відношенню до вісі симетрії.

На практиці ця проблема має набувати такої форми. Нехай маємо таке сполучення $\{1{,}1{,}1{,}1{,}0{,}0\}$. Тобто чотири моделі дали правильний результат та дві неправильний. Згідно з принципом голосування простою більшістю результат ансамблю буде правильним. Однак існує множина даних, яка симетрична відносно вісі S та відповідає сполученню $\{0{,}0{,}0{,}0{,}1{,}1\}$. Для цієї множини підхід простої більшості призведе до невірного рішення ансамблю. Необхідним підходом до цієї множини буде зворотній принцип – простої меншості. Таким чином, визначається

проблема розрізнення множин симетричних сполучень.

Однак ця проблема відсутня для множин симетричних граничних сполучень, а саме $\{1,1,1\}$ та $\{0,0,0\}$. При достатній кількості моделей в ансамблі множина даних, які неправильно класифікували усі моделі, вироджується і є порожньою. Отже, є можливість однозначно визначити дані, які визначаються за правилом and-gate, тобто всі моделі дали однакове рішення. У випадку виродженості множини неправильних рішень, усі моделі дали однакове рішення і воно є правильним для цієї множини даних. Це і формує сутність інформаційного ядра класифікації.

Формування набору сполучень рішень моделей

Для одержання даних фактичного розподілу рішень та формування набору сполучень було використано бібліотеку Scikit-Learn та корпус, який широко використовується Reuters-21578. Для отримання рішень не використовувався весь корпус, а було використано обмежену кількість даних. Загальна кількість даних 1000, з них для тестування – 134, які класифікуються за 10-ма найбільш розповсюдженими класами.

Масив класифікаторів, за яких множина усіх однакових неправильних класифікацій є виродженою, містить три класифікатора SVN, NB та перцептрон. При цьому формується такий розподіл сполучень результатів (рис. 2).

Таким чином, усі дані набору для тестування були правильно класифіковані хоча б одним класифікатором з множини ансамблю.

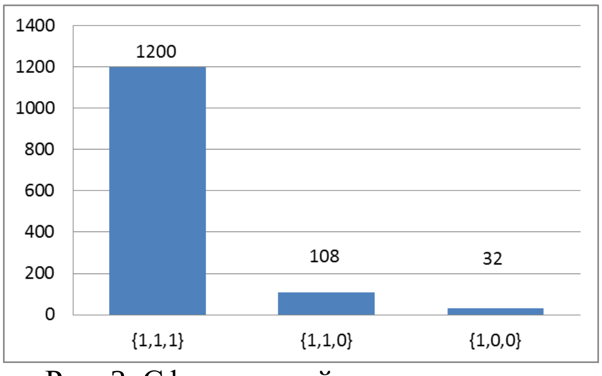


Рис. 2. Сформований ряд сполучень

Висновок

Проведені дослідження у напрямку формування множини ансамблю класифікації дозволили виявити підходи, які дають змогу визначати вплив на перерозподіл даних внаслідок розширення множини. Це дало змогу отримати та практично перевірити наступні результати.

1. Встановлено що розширення множини ансамблю класифікаторів формує набір множин сполучень рішень моделей та призводить до перерозподілу даних сусідніх множин сполучень.
2. Перерозподіл даних дозволяє сформувати множину ансамблю, завдяки якій вироджується множина неправильних рішень усіх класифікаторів.
3. Визначено умови формування множини класифікаторів та встановлені критерії її достатності для вирішення завдання розпізнання.
4. Встановлено, що подальше розширення множини класифікаторів понад критерії достатності привносить помилки класифікації до множини правильних рішень усіх класифікаторів.
5. Досягнення умов достатності множини ансамблю дозволяє практично застосувати принцип агрегатування за схемою and-gate.
6. Досягнення умов достатності множини ансамблю забезпечує формування множини спільних однакових рішень класифікаторів, і це рішення, яке спільне для усіх класифікаторів, буде правильним.
7. Розглянуто принципи агрегатування та їх застосовність до ансамблю класифікаторів.
8. Проведено аналіз практичної застосовності принципів формування та переваг агрегатування та формування ансамблів.

Використання принципів повноти ансамблю та критеріїв достатності дозволяє сформувати ефективні правила агрегатування множини класифікаторів.

Література

1. Mueller, D. (2003). Public Choice III. Cambridge: University of Cambridge Press.
2. Gerecke, U., Sharkey, N.E., Sharkey, A.J. (2003). Common evidence vectors for self-

organized ensemble localization, Neurocomputing, 55(3), P. 499-519.
3. Schapire, R.E., (1990). The strength of weak learnability, Machine learning, 5(2), P. 197-227.
4. Zhu, M (2015). Use of majority votes in statistical learning. Wiley Interdisciplinary Reviews: Computational Statistics, 7(6), P. 357 - 371.
5. Vardeman, S.B., Morris, M.D. (2013), "Majority Voting by Independent Classifiers can Increase Error Rates," The American Statistician, 67, P. 94-96
6. Vardeman, S.B., Morris, M.D. (2013), "Majority Voting by Independent Classifiers can Increase Error Rates," The American Statistician, 67, 94–96: Comment by Stuart Baker, Jian-Lun Xu, Ping Hu and Peng Huang and Reply The American Statistician, 2014, 68(2), P. 125-126
7. Манзюк Э.А., Бармак А.В., Крак Ю.В., Касьянюк В.С. (2018) Определение информационного ядра при классификации документов. Проблемы управления и информатики, 2, 78-86. DOI: 10.1615/JAutomatInfScien.v50.i4.30

References

1. Mueller, D. (2003). Public Choice III. Cambridge: University of Cambridge Press.
2. Gerecke, U., Sharkey, N. E., Sharkey, A. J. (2003). Common evidence vectors for self-organized ensemble localization, Neurocomputing, 55(3), P. 499-519.
3. Schapire, R.E., (1990). The strength of weak learnability, Machine learning, 5(2), P. 197-227.
4. Zhu, M (2015). Use of majority votes in statistical learning. Wiley Interdisciplinary Reviews: Computational Statistics, 7(6), P. 357-371.
5. Vardeman, S.B., Morris, M.D. (2013), "Majority Voting by Independent Classifiers can Increase Error Rates," The American Statistician, 67, P. 94-96
6. Vardeman, S.B., Morris, M.D. (2013), "Majority Voting by Independent Classifiers can Increase Error Rates," The American Statistician, 67, P. 94–96: Comment by Stuart Baker, Jian-Lun Xu, Ping Hu and Peng Huang and Reply The American Statistician, 2014, 68(2), P. 125-126
7. Manziuk, E.A., Barmak, A.V., Krak, Y.V., Kasianiuk, V.S. (2018). Opredelenie informatsionnogo yadra pri klassifikatsii dokumentov. Problemyi upravleniya i informatiki, 50(4), P. 25-34. DOI: 10.1615/JAutomatInfScien.v50.i4.30

RESUME

O.V. Barmak, E.A. Manziuk, Iu.V. Krak, A.I. Kulias

Principles and approaches to the formation of ansambles based on analysis results

The article considers principles of the formation and application of classifiers' ensembles. Particular attention is paid to the application of the basic foundations of the creation of ensembles. The advantages of using the approaches of aggregation of classifier solutions are considered. An analysis of the principles of data redistribution with the extended ensemble made it possible to formulate requirements for the classifiers of the ensemble. On the basis of what was found the conditions for the selection of classifiers and the criteria for their adequacy. On the basis of the sufficiency criteria, the applicability conditions for the principle of aggregation for boundary sets were established. An analysis of the ambiguity of decision making for symmetric sets of classifiers is carried out. The conditions for forming a set of classifiers and the criteria for its adequacy for solving the recognition problem are determined. It is established that further expansion of the set of classifiers over the sufficiency criterion brings classification errors to the set of correct solutions of all classifiers. The extension of the classifier set allows us to form a set of connections that is a Pascal triangle and to analyze the marginal redistribution of data in the process of increasing the ensemble.

The conducted researches in the direction of forming the set of ensemble of classification allowed to reveal approaches that allow to determine the effect on data redistribution as a result of the expansion of the set. This made it possible to get and actually test the results. It is established that the expansion of the set of ensemble of classifiers forms a set of sets of combinations of model decisions and leads to the redistribution of data of neighboring sets of combinations. Redistribution of data allows us to form a set of ensembles, through which the set of invalid moves degrades all classifiers. The conditions of formation of a set of classifiers are determined and the criteria of its sufficiency for solving the problem of recognition are established. It was established that further expansion of the set of classifiers over the sufficiency criterion brings classification errors to the set of correct solutions of all classifiers.

Надійшла до редакції 3.10.2018